National Center for Biotechnology Information

National Library of Medicine

Summary of Research 1992-1993

The Basic Research Branch and the Information Engineering Branch at the NCBI are comprised of a multidisciplinary group of scientists who carry out research on fundamental biomedical questions at the molecular level by developing and utilizing mathematical, statistical and other computational methods. The approach includes both the theory and the technology of the methods themselves. These two lines of research are mutually reinforcing and complementary. The basic research has led to new practical methods and the use of methodology has opened new areas of research.

Several analytical algorithms and methods have been developed to complement the investigation of biomedical problems. Analytical methods and algorithms have focussed on sequence similarity, structural modelling and genome analysis. Pattern matching and sampling methods for biopolymer sequence data, which include the statistics of sequence comparisons, have been applied to investigate such protein families as the ras-like GTPases, steroid receptors, cold shock domain proteins, HMG-1 box containing proteins, and NTPases, as well as other analyses of different sequence motifs. Biomolecular structural investigations include such projects as the analysis of packing contacts in protein crystals, 2D lattice models of proteins, 3D modelling of ribonucleic acids, protein threading and the energy distribution of compact states of a peptide. Genome analyses have been a significant part of both in the method development and analysis of several different genomes including that of *E. coli,* RNA viruses and humans.

In order to attempt to unify biotechnology information, explicit descriptions for biosequence objects have been defined using an international standard data description language, ASN.1. Innovative approaches have been taken for database design to permit the integration and linkage of vast amounts of sequence-related data. ASN.1 software libraries have enabled the rapid development of software tools for the retrieval and analysis of information from these databases.

Text information retrieval methods and document analysis has been an important component of the basic research. New techniques in use and under evaluation for the analysis of large amounts of text include Gibbs sampling approaches, Bayesian methods and the building of neighboring lists.

NCBI Intramural Research Projects

Structure/Function:

Steroid responsiveness and steroid receptor binding sites in the HIV-1 LTR.

Proteins that interact with ras-like GTPases.

Information and molecular recognition in DNA-protein interactions.

Identification and mapping of the human homolog of the yeast cell cycle gene.

Subtle sequence patterns in DNA-binding complexes.

Computer analysis of low-complexity amino acid sequences.

Molecular novelty and conservation in bacterial protein sequences.

The identification of fossil sequences by exhaustive comparison of protein sequences from evolutionary distant organisms.

Analysis of conserved amino acid sequence motifs in NTPases.

The cold shock domain (CSD) protein motif.

The HMG-1 box protein motif.

DNA sequence complexity and mutational dynamics.

Biomolecular Structure:

Analysis of packing contacts in protein crystals.

Extended two dimensional lattice models of proteins.

3-D computer modeling of ribonucleic acids.

Structure prediction by protein threading.

The energy distribution of the compact states of a peptide.

Theory of Sequence Analysis:

The statistics of sequence comparison.

A depth-first search algorithm for detecting patterns in protein sequences.

Gibbs sampling methods for the analysis of biopolymer sequence data.

Analysis of reliability of molecular sequence data.

Hardware Design:

Bioscan - a VLSI-based system for biosequence analysis.

Software and Database Design:

Unification of biotechnology information.

Portable toolkit for scientific software.

Design and uses of a public resource of expressed sequence tags.

Development of a transcription factors database.

Databases for molecular modeling.

A representative set of protein sequences for similarity.

Text Retrieval and Document Analysis:

Textual information retrieval testing.

Automatic Bayesian methods in text retrieval.

A document processing system.

Strategies for finding relevant documents among neighbor lists.

Situation theory as a model for ontological engineering and knowledge.

Gibbs sampling as an approach to document retrieval.

Genome Analysis:

Informatics analysis of the *E. coli* genome.

Genome organization and evolution of RNA viruses.

Sequence analysis methods adapted to large scale sequencing projects.

Comprehensive computer analysis of *E. coli* genes.

Identification of the Kallmann syndrome gene (xp22.3) and of new genes in the HLA class III region (6p21.3).

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00001-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Steroid responsiveness and steroid receptor binding sites in the HIV-1 LTR

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and Institute affiliation)*

David Ghosh, Staff Fellow, NCBI

COOPERATING UNITS *(If any)*

Laboratory of Viral and Molecular Pathogenesis, NINDS, NIH (E. Verdin)

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.2	0.2	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☒ (b) Human tissues ☐ (c) Neither

☐ (a1) Minors

☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

Steroids have some degree of clinical relevance in HIV pathogenesis since (a) infected patients frequently show altered plasma cortisol levels and (b) steroids are used in the treatment of certain AIDS-associated opportunistic infections. A recently mapped glucocorticoid receptor binding site has been tested for function in transient expression assays. The preliminary results from these studies, which are similar to observations reported by other researchers, suggest that this binding site is a weak but complex regulatory element.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-00004-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Development of Transcription Factors Database

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

David Ghosh, Staff Fellow, NCBI

COOPERATING UNITS *(if any)*

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.3	0.3	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☑ (c) Neither

☐ (a1) Minors

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

During this project period, the size of this database increased by 14%. A few minor alterations to the design of the tables have been made and querying methods for the generation of sites maps have been tested. TFD has now been implemented in an object-oriented database management system, permitting the testing of different types of data structures as solutions to the problem of computational representation of complex biological entities. The 7.0 release of the database contains 2106, 1016, 523, 1626, 2155, 38, 2876, 7391, 5757 records in the clones, domains, factors, polypeptides, sites, methods, n_pointers, references, x_pointers tables.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00005-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Information and Molecular Recognition in DNA-protein interactions

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

David Ghosh, Staff Fellow, NCBI
W.J. Leonard, Laboratory of Pulmonary Immunology, NHLBI, NIH

COOPERATING UNITS *(if any)*

Laboratory of Pulmonary Immunology, NHLBI, NIH (W.J. Leonard);

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.5	0.5	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

A recently reported sequence motif is present in two different classes of DNA binding domains, the NRD (Rel family) domain and the zinc finger, both of which bind NF-kappa-B sites. This sequence motif was tested in the context of the zinc finger by mutagenesis of the corresponding residues. These mutagenesis studies indicate that these residues in the zinc finger play some direct or indirect role in sequence recognition.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00009-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Analysis of Packing Contacts in Protein Crystals

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Stephen H. Bryant, Senior Investigator, BRB, NCBI, NLM
J.A. Bell, Assistant Professor, Rensselaer Polytechnic Institute
S. Dasgupta, Graduate Student, Rensselaer Polytechnic Institute

COOPERATING UNITS *(if any)*

Chemistry Department, Rensselaer Polytechnic Institute, Troy, NY (J.A. Bell, S. Dasgupta)

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.05	0.05	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

The goal of this project is to identify common features of packing interactions in protein crystals, by statistical survey of structures in the Brookhaven Protein Data Bank. We have used functions in the PKB suite to identify a non-redundant subset of structures in the Protein Data Bank which have valid crystal symmetry information. For these structure we have tabulated crystal-packing contacts by residue and residue-pair type, and conducted statistical analyses. An interesting feature identified is the preferential participation of amino acids with small side chains, suggesting that lattice packing in diffraction-quality crystals must involve rigid intra-molecular contact of the polypeptide backbone. Another interesting feature is a clear correlation between the density of packing contacts and the resolution of available diffraction data, suggesting that lattice packing can reduce crystal disorder. The significance of this project is in providing a summary of protein-protein interactions as observed in crystals. These are very different from intra-molecular or oligomer-docking contacts, and they may provide a model for interactions expected in loosely-associated biological complexes.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00010-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Sequence analysis methods adapted to large scale sequencing projects

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Jean-Michel Claverie, Visiting Scientist, NLM, NCBI

COOPERATING UNITS *(if any)*

None

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.33	0.33	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither

☐ (a1) Minors

☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

The biological interpretation of the data generated by large scale sequencing projects is plagued by unmanageable large program outputs the size and the noise/signal ratio of which obscure the truly relevant findings. This problem can be attributed to various causes: the increasing size of the databases, their quality (redundancy, experimental and clerical errors) and some intrinsic properties of biological sequences such as repeated elements and local compositional bias. In order to alleviate these problems we have developped a set of programs that are now routinely used in an information enhancement step prior to the analysis of large body of sequence data. Following this steps, the output of gene identification and sequence comparison programs become biologically and statistically interpretable without further processing. In addition we have developped a suite of independent modules that can be used in sequence to automatically analyze large body of experimental data. The power of these tools has been demonstrated in the context of collaborations with experimental groups generating a large amount of sequences (see project Z01-LM-00011-01-BRB). In the meantime, we are also developping new sequence analysis methods than can be both applied to the quality control of sequences already in the databases or to the interpretation of newly determined ones.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00011-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Identification of the Kallmann's syndrome gene (Xp22.3) and of new genes in the HLA class III region (6p21.3).

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Jean-Michel Claverie, Visiting Scientist, NLM, NCBI
Daniel Cohen (director, CEPH/Genethon, Paris, France)
Christine Petit (head of laboratory, Institut Pasteur, France)

COOPERATING UNITS *(if any)*

CEPH/Genethon, Paris, France
Institut Pasteur, Paris, France

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.33	0.33	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

Two large human genomic sequence contigs have been analyzed with the set of programs (BLAST, XNU, Xblast, EXON, etc) and databanks (NRDB, Geninfo) uniquely available at NCBI. Those studies have lead to the identification of the gene for Kallmann's syndrome (ADML-X, Xp22.3) and of a putative new transcription factor in the HLA class III region (6p21.3).

1) Identification of the Kallmann's syndrome gene

X-linked Kallmann's syndrome (affecting 1/10000 male) is a defined by simultaneous defect in the migration of GN-RH neurones and in the fasciculation of the olfactory neurones during the early development of the brain. Molecular genetics analysis led to the sequencing (Institut Pasteur, Genethon) of a 67kb contig in which two exons totalling lesss than 450 nucleotides where identified by a combination of computer analysis techniques. The study of the protein sequence encoded by a full lenght cDNA revealed an original assembly of fibronectin and anti-protease domains, consistent with its role as a neurone targeting molecule.

2) Analysis of a 90kb contig in the HLA class 3 region

The same combination of computer technique was applied to the analysis of a 90 kb sequence contig (CEPH/Genethon). This region revealed an unusually dense clustering of Alu , the genomic structure of the Bat2 gene (25 exons) and a putative new member of the NFkB family of transcription factors.
This project will be complete after publication of the results.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00012-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

The identification of fossil sequences by exhaustive comparison of protein sequences from evolutionary distant organisms.

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Jean-Michel Claverie, Visiting Scientist, NCBI, NLM
David J. Lipman, Director, NCBI, NLM
Phil Green, Washington University, St. Louis, MO

COOPERATING UNITS *(if any)*

Genetics Department, Washington University, St. Louis

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.33	0.33	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

Sequence alignments of homologue proteins from evolutionary distant organisms are used to pinpoint regions of structural and functional importance. Over long periods only the most constrained segments retain a detectable similarity with each other. This concept was extended to the whole database, by performing cross-comparisons of comprehensive sets of sequences from various kingdoms and phyla with evolutionary distances ranging from 2 billion years for the eukaryote/eubacteria divergence to 550 million years for the coelomate radiation. Significant similarities between these sets thus correspond to strongly conserved ancestral features. Using a series of matching/orthogonalization procedures, 500 independent ancestral types were detected within contemporary sequences. This fossil set only represents 4% of the original database but significantly matches 40 % of the whole. Thus, it realizes a 10-fold enrichment in sequences of the greatest structural/ functional significance and is an optimal source for the definition of motifs. Approximately 200 of those highly conserved sequences correspond to proteins the role of which is not obviously central, and warrant further analysis. Theoretical computations suggest that the 500 ancestral types defined so far constitute most of the fossil sequences detectable in modern sequences. This is consistent with another independent study comparing 3 large new datasets: partial cDNAs from human and nematode and ORFs from chromosome III of yeast. Thus, known proteins might already include representative for most ancestral features antedating the coelomate radiation. A database of 550 ancient sequence prototypes has been constituted and made available to the community by computer network.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00013-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

BioSCAN - A VLSI-Based System for BioSequence Analysis

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Stephen Altschul, Senior Staff Fellow, NCBI, NLM
Warren Gish, Staff Fellow, NCBI, NLM
C. Thomas White, Molec. Biol. & Biotech. Prog., UNC, Chapel Hill, NC
Raj Singh, Dept of Computer Science, UNC, Chapel Hill, NC
Peter B. Reintjes, Quintus Computer Systems, Mountain View, CA
Jordan Lampe, Dept. of Comp. Sci. & Eng., University of Washington
Bruce W. Erickson, Dept. of Chem., UNC, Chapel Hill, NC
Wayne D. Dettloff, MCNC, Research Triangle Park, NC
Vernon L. Chi, Dept. of Comp. Sci., UNC, Chapel Hill, NC
S. Tell, Dept. of Comp. Sci., UNC, Chapel Hill, NC
D. Hoffman, Dept. of Comp. Sci., UNC, Chapel Hill, NC

COOPERATING UNITS *(if any)*

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.1	0.1	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

The BioSCAN project involves the design and construction of a computer system for very rapid comparison of protein and DNA sequences, built abound a special-purpose VLSI chip designed and manufactured specifically for the project. The work has been conducted primarily by the Computer Science Department of the University of North Carolina at Chapel Hill under a grant from the NSF. However, I have been involved as a consultant on the design and use of the system. This past year, has seen the development of a new board that uses two BioSCAN chips and is hosted by a PC. Demonstration software had been developed and is being tested and refined.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00014-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

The Statistics of Sequence Comparison

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Stephen Altschul, Senior Staff Fellow, NCBI, NLM
Warren Gish, Staff Fellow, NCBI, NLM
Samuel Karlin, Dept. of Mathematics, Stanford University, Stanford, CA

COOPERATING UNITS *(if any)*

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.5	0.5	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

This project is a continuing study of questions concerning what similarities can be expected to occur purely by chance when two protein or DNA sequences are compared. A subsidiary and related question concerns the definition of scoring systems that are optimal for distinguishing biologically meaningful patterns from chance similarities. Advances this year include: a) The publication of a scoring system for molecular sequence comparison that is sensitive to similarities at all evolutionary distances, including an analysis of its statistics: This work was completed mainly in the previous year, but was published this year. It details how a single "amino acid substitution matrix" is best adapted to detecting similarities at a single evolutionary distance, and describes how multiple matrices may be used to cover the complete range of detectable similarities. The statistics of this multiple matrix comparison method are studied (Altschul, 1993). b) Statistics for the sum of the scores of high-scoring segment pairs: In collaboration with Samuel Karlin, I have described the statistical behavior of Sr, the sum of the scores of the r highest-scoring distinct segment pairs (Karlin & Altschul, 1993). These statistics are the first rigorous approach to the statistics of scored alignments with gaps. A program to calculate the distribution of Sr, involving a double integral, has been developed with the assistance of Warren Gish and John Spouge. c) The development of Poisson and sum statistics for consistent high-scoring segment pairs:Comparison of protein of DNA sequences frequently yields multiple high-scoring segment pairs. A combined assessment of these segment pairs generally is appropriate only when they may be combined, with the introduction of gaps, into a single consistent alignment. This requires a modification of the sum statistics just described, and of the Poisson probability for finding at least distinct segment pairs with score at least S. The imposition of consistency at once weeds out many "chance" alignments, and increases the reported significance of the true ones. The statistics of consistent segment pairs have now been described (Karlin & Altschul, 1993), and they have be incorporated into the BLAST programs

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00015-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Design and uses of a public resource of expressed sequence tags.

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Mark S. Boguski, Senior Medical Staff Fellow, BRB, NCBI, NLM
Carolyn Tolstoshev, Visiting Associate, IEB, NCBI, NLM

COOPERATING UNITS *(if any)*

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.4	0.4	0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

Expressed Sequence Tags (ESTs) are partial nucleotide sequences derived from clones that are randomly-selected from cDNA libraries. The accumulation and analysis of ESTs has become an important component of genome research. The rate of EST sequence acquisition is accelerating and >25,000 ESTs have been accessioned into our database, dbEST, during the past two years. dbEST aims to provide value-added annotation and timely access to new data and analyses.

We have developed a relational database to manage EST data as well as a custom software system for data analysis. This system performs periodic homology updates after screening the query sequences and masking contanimating or uninformative subsequences. Also stored in the database is information about the availability of physical DNA clones and genetics map locations. These data and analyses are made available to the public in four ways: 1) dbEST is a line-item database for network and email BLAST searches; 2) full reports are available from est_report@ncbi.nlm.nih.gov; 3) as a FASTA-formatted file for anonymous ftp; and 4) in the new EST Division of GenBank. Plans are underway to implement an Internet Gopher service for EST information retrieval. We will also be expanding our storage and retrieval capability for genetic mapping data and will accept submissions of exon-trapped sequences as well as ESTs.

A number of discoveries of medical significance have already been made using the dbEST resource to accelerate the cloning of human genes whose homologs have already been characterized in other species.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00016-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Proteins that Interact with Ras-like GTPases.

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Mark S. Boguski, Senior Medical Staff Fellow, BRB, NCBI, NLM
Francis Collins, Howard Hughes Investigator, Univ. of Michigan, Ann Arbor, MI
Frank McCormick, Vice President of Research, Onyx Pharmaceuticals, Emeryville, CA
Scott Powers, Assist. Prof., Robert Wood Johnson, Medical School, Piscataway, NJ
Michael Wigler, Senior Scientist, Cold Spring Harbor Lab., Cold Spring Harbor, NY

COOPERATING UNITS *(If any)*

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.4	0.4	0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

ras, rab and rho/rac are small membrane-bound GTPases that function as key control switches in mitogenic signal transduction, membrane trafficking and cytoskeletal organization, respectively. Oncogenic mutants of ras genes are the transforming genes most frequently found in human cancers and ras has been one of the most intensively studied proteins of the past decade. It is now becoming clear that an expanding menagerie of other proteins interact with ras and influence its GTPase and guanine nucleotide exchange activities. For ras to function (or mal-function as the case may be), it must be located at its site of action on the inner surface of the plasma membrane. ras is targeted to and anchored to the membrane by virtue of an isoprenoid lipid group that is post-translationally attached to ras by a heterodimeric enzyme called farnesyltransferase (FTase). Other ras-like proteins are modified by geranylgeranylation (type I and II enzymes). Inhibitors of FTase have recently been developed and show considerable promise as a new class of anticancer drugs.

Through database searching and multiple sequence alignment, we have identified highly-conserved domains and sequence motifs in GTPase-activating proteins (GAP), guanine nucleotide-releasing proteins (GNRP), guanine nucleotide dissociation inhibitors (GDI), and prenyltransferase subunits. We discovered pleckstrin homology (PH) domains in p120 rasGAP and rasGNRP. Sequence homologies between the human choroideremia gene product, rabGDI and the yeast protein Mrs6 were identified and characterized and the latter inferred to be "rab escort protein" that forms the third subunit of type II geranylgeranyltransferase.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER: Z01-LM-00020-02-BRB

PERIOD COVERED
October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*
Textual Information Retrieval Testing

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*
W. John Wilbur, Senior Scientist, BRB, NCBI, NLM

COOPERATING UNITS *(if any)*

LAB/BRANCH
Basic Research Branch

SECTION

INSTITUTE AND LOCATION
NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.3	0.3	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☑ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

New approaches to testing the effectiveness of retrieval methods have been studied. A method of testing retrieval performance which involves the comparison of statistically independent retrieval methods has been developed. A second method of testing retrieval based on modelling the document collection and the relevance relation has been investigated and compared with the previous method. This second method involves the hypergeometic probability distribution and yields results quite consistent with the first. A paper has been published describing this modelling method.

A new measure of retrieval performance based on information theory has been discovered. This measure is simple to apply and calculates the number of bits of information produced by a ranked retrieval method in concentrating relevant documents in the first n ranks in the retrieval operation. It has many intuitively desirable properties and agrees with precision-recall curves when the latter allow the unambiguous comparison of different methods of retrieval. A paper has been published describing the methodology.

Studies have been performed to evaluate the sensitivity of retrieval testing to the number of queries in a test set and to the particular measure used. Several of the classical test sets (CRAN, CISI, and MED, etc.) as well as a test subset of Medline in the area of molecular biology have been studied. Special attention has been paid to the problem of large databases where only incomplete sampling is possible.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00021-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Automatic Bayesian Methods in Text Retrieval

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

W. John Wilbur, Senior Scientist, BRB, NCBI, NLM

COOPERATING UNITS *(if any)*

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.3	0.3	0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither

☐ (a1) Minors

☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

An automatic approach to applying Bayesian methods in text retrieval has been developed. This is a form of relevance weighting of search terms but departs from the usual approach in two ways which complement each other. First, the usual approach involves the assignment of relevance weights to the search terms in a single query based on the documents that are and those that are not relevant to the query. This involves generally a small number of relevant documents and hence a statistical sample that is difficult to use in making any globally significant inferences about the value of the terms involved. We modify the usual approach by taking the average of the importance of a term over all the queries in which it occurs. We study the case when the set of queries is the set of documents so that the global term relevance weight is a well defined concept. Second, the usual approach is limited to the case when one has human judgments of the relevance of documents to queries. This has limited the use of the method to certain test sets where the relevance relation is known or to relevance feedback situations. Our approach is to replace the relevance relation by the relation of high scoring pairs of query and document using the vector cosine method of retrieval. Because the latter is an automatic method we are able to generate the required statistics in an automatic manner. While this latter approach will undoubtedly have more error than human relevance judgments the larger sample size involved in global weighting helps to offset this problem. Local weighting is introduced in an ad hoc manner and the resultant retrieval is found to be somewhat superior to vector cosine retrieval.

Their are two problems with the model just described. First it does not incorporate local term weighting in a natural Bayesian manner and second it does not provide a correction for document length. We have developed a new model based on cluster concepts that remedies these two problems while allowing the model to remain completely Bayesian. This performs at the same basic level as the one already described in which local weights are treated ad hoc. It does however allow one to see the actual log odds predictions of relevance. These exceed the observed log odds of relevance by 13:1 which gives an interesting perspective on term dependency.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER: Z01-LM-00022-02-BRB

PERIOD COVERED
October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*
A Document Processing System

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*
W. John Wilbur, Senior Scientist, BRB, NCBI, NLM

COOPERATING UNITS *(if any)*

LAB/BRANCH
Basic Research Branch

SECTION

INSTITUTE AND LOCATION
NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.2	0.2	0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☑ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

A system of software has been developed for the purpose of finding the closely related documents in Medline. The system consists of over thirty-five programs written in the C language plus a number of utility programs. The system has a number of unique features:

1) It is highly modular so that alterations in the system are relatively simple to perform.
2) The system currently operates on Medline data in the ASN1 format but a change in the interface portion of the system would allow it to be applied to any large database consisting of discrete textual records.
3) The system is designed with a degree of security against loss of data due to operating system crashes or power outages.
4) All data processed by the system is stored in permanent form as inverted file structures, etc. These structures are updateable so that new data may be continually added to the system as it becomes available.
5) Documents are compared with each other using a Bayesian form of analysis and the statistics on which the relevance weighting of terms is based are derived from previous document comparisons. These statistics are updated with each new cycle of processing.
6) The probability that documents are related is computed by the system based on a scaling of the raw scores produced using a set of document pairs that have been judged for relatedness by human judges. This scale is recalculated each time term weights are updated and it is calculated differently for documents with as opposed to documents without abstracts.

An analysis of the most glaring failures of the system, as identified on the test set used for scaling of document similarity, has been carried out. This shows that a significant part of the problems experienced may be due to the common occurrence of several areas with different levels of description and different levels of uniformity in description in a single document. The numerical representation of these descriptive areas within a document does not in general correlate with their importance in defining document content.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00023-02-BRB

PERIOD COVERED

October 1, 1992 to June 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Strategies for Finding Relevant Documents among Neighbor Lists

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

W. John Wilbur, Senior Scientist, BRB, NCBI, NLM
Leona Coffee, Library Associate, NLM, NIH

COOPERATING UNITS *(if any)*

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.2	0.2	0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☑ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

One method for gaining access to a database is for a searcher to choose some key terms or write a short description of the topic of his interest. The resultant key terms which are given directly or processed out of the written request are then used to make a vector for searching the database for related documents. It is well known that the specificity of such a search is not as good as that which would be obtained from a more definitive description of the searchers interest, a description which he in many cases can only give after he has found an article of interest. This is the basis for so-called relevance feedback procedures. Relevance feedback makes use of relevance judgments made on already retrieved material to focus a search. In this study we are concerned with a special kind of relevance feedback.

We deal with databases in which for each document the list of the top documents in relation to the given document have already been computed. Our purpose is to discover the most efficient use of these precomputed neighbor lists to facilitate a search which may begin by a search based on several key terms as described in the previous paragraph. When the first relevant document is found it comes with a precomputed list of neighbors. The question is whether to look back for the next relevant document on the initial list or to look on the list of precomputed neighbors associated with a relevant document. After the second relevant document has been identified then one again has the question of how to use its neighbor list optimally and so on. We have tried a number of different strategies and have found that there is definite improvement in using the neighbor lists and have quantified the performance obtained on the CISI, CRAN, CACM, and MED test sets of documents. A paper describing these results is in press.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00025-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Computer analysis of low-complexity amino acid sequences

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

John Wootton, Visiting Scientist, BRB, NCBI, NLM
Scott Federhen, Computer Scientist, NCBI, NLM

COOPERATING UNITS *(if any)*

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
1.1	1.1	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☑ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

The goal of this project is to define, classify and analyze, using computational analysis, all segments of protein sequences of improbably low compositional complexity. These include residue clusters of predominantly one or a few amino acid types, which commonly contain homopolymeric tracts or mosaics of these, aperiodic patterns and sections of low-period repeats. The abundance of these segments in sequence databases has been determined and their properties are being related to evidence of biological functions and protein structure, dynamics and assembly.

A. Methods: Different formal definitions of local compositional complexity were used to make unbiassed identification of low-complexity segments, irrespective of their specific residue clustering or repeat patterns, at different levels of stringency. Algorithms were refined to (a) select segments for further study and (b) filter out non-informative segments prior to database searches. New methods for automated classification and neighboring of low-complexity sequences have been developed.

B. Abundance and biological properties: Approximately 15% of the residues in protein databases are in low-complexity segments of typically 15-50 amino acids, and approximately 55% of proteins contain one or more such segments. This fraction has increased significantly in the last year, reflecting the greater number of database entries from genomic sequences determined without regard to function. Interspersed low-complexity sequences are particularly abundant in many eukaryotic proteins crucial in morphogenesis and embryonic development, transcriptional regulation, signal transduction and aspects of cellular and extracellular structural integrity and interactions.

Significance of project: The project has highlighted the high abundance and biological importance of low-complexity protein segments and emphasized the relative lack of knowledge of their molecular structure and dynamics. Low complexity segemnts evidently have polymorphic, non-compact structures and dynamics which are necessary for biological function. The new computer methods are valuable in elminating many artefacts in sequence database searches and alignment analysis.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00026-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Molecular novelty and conservation in bacterial protein sequences

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

John Wootton, Visiting Scientist, BRB, NCBI, NLM

COOPERATING UNITS *(if any)*

Experimental Laboratories at Universities of Leeds and Sussex, England, University of Southern Illinois, Carbondale, John Innes Institute, England and University of Lyon, France.

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.2	0.2	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☑ (c) Neither

☐ (a1) Minors

☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

Protein sequences deduced from gene sequences of diverse bacteria and archaebacteria are yielding a wealth of new knowledge on protein functions, interactions and evolution. Some novel findings were studied by concerted methods including directed mutagenesis, spectroscopic/enzymological analyses and computer analyses of sequence databases. Findings include:-

A. Bacterial homologs of major "eukaryotic" protein families. Bacterial homologs of Serine/Threonine protein kinases were further investigated in cyanobacteria and archaebacteria. These are evidently relatively similar to the protein kinase C subfamily, but their phosphorylation specificity remains to be confirmed.

B. Novel proton exchange mechanism in glutamate dehydrogenase. This enzyme differs from other NAD/NADP dependent dehydrogenases in lacking a catalytic histidine residue. From sequence conservation and site-directed mutagenesis of the E. coli enzyme, in comparison with the crystal structure of the Clostridium symbiosum enzyme, a triad of lysine residues was identified in the active site. These have novel protonation properties and evidently act in concert in dicarboxylate substrate binding, deprotonation of ammonium and other proton exchange steps of glutamate dehydrogenase catalysis.

C. Sequence families in complex bacteria. New sequences from multicellular and differentiating bacteria, Streptomyces, Myxococcus and various cyanobacteria and archaebacteria, were investigated by global database sequence similarity searches. More than 50 percent of the classifiable protein sequences did not have counterparts in E. coli and other well-studied enteric bacteria, and many of these had eukaryotic homologs. Examples of low-complexity segments and multiple repeats are emerging with increasing frequency.

Significance of project: Genome sequences from metabolically and morphogenetically diverse bacteria continue to provide a rich and cost-effective source of new discoveries on the molecular functions and evolution of proteins.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00030-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Analysis of the Reliability of Molecular Sequence Data.

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Warren Gish, Staff Fellow BRB, NCBI, NLM

COOPERATING UNITS *(if any)*

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
.05	.05	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither

☐ (a1) Minors

☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

DNA sequence is error prone and the presence of sequence data errors can reduce the sensitivity mo of database searches, particularly for distantly related homologs.

The reliability of one class of sequence search algorithm BLAST has been explored operating on sequence data with different levels of error introduced artificially. In addition, a version of this algorithm, BLASTX which translates nucleic acid sequences in all possible reading frames and search these conceptually translated protein sequences against protein sequence databases has been used to evaluate search performance based on raw cDNA sequence which is now being deposited in molecular sequence databases as expressed sequence tag (EST) sequence data. The use of codon utilization information has also been incorporated into the BLASTX algorithm to identify coding regions through a combination of sequence alignment and codon utilization information. This makes the identification of coding regions more robust to the presence of data errors in the query or database.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00032-02-IEB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Portable Toolkit for Scientific Software

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

James M. Ostell, Chief, IEB, NCBI, NLM
Warren Gish, Staff Fellow, NCBI, NLM
Jonathan Kans, Staff Fellow, NCBI, NLM
Gregory Schuler, Staff Fellow, NCBI, NLM

COOPERATING UNITS *(If any)*

LAB/BRANCH

Information Engineering Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
4	4	0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

Objectives: To determine if a single environment can be used to develop scientific software which operates without change on all major hardware/software platforms in use by NIH supported scientists. **Methods employed:** Design and research on a layered system of software. Development of real production quality tools to test the robustness of design and implementation in the real world. **Major findings:** It is possible to write a single scientific "C" language program which will run without change on Macintosh, IBM PC class, Microsoft Windows, Sun UNIX, Silicon Graphics UNIX, DEC ULTRIX, VAX VMS, and IBM 3090 AIX computers. Such programs can not only perform routine computation, but use various modern windowing systems (Macintosh, MS-Windows, X11 Motif) for user interface and International Standards Organization (ISO) protocols for structured data exchange. A system of software layers represented by "C" language libraries provides a flexible robust environment. The CoreLib layer is a thin interface between a single logical view of program flow and operations (on the programmer side)and the details of implementing that view on the various supported platforms. The AsnLib layer, built on the CoreLib layer, provides portable utilities for structured, standardized data exchange using Abstract Syntax Notation 1 (ISO 8824, 8825). The Object layer, built on the AsnLib layer, reads and writes ASN.1 formatted data streams in and out of "C" structures available to the programmer. The Vibrant layer, built on the CoreLib layer, provides a single programming interface to three different windowing systems. Vibrant supports both a very simple view of user interaction typical of scientific programs and the very complex view of modern highly interactive visual interfaces. A variety of production quality software products have been developed using the Toolkit, and are in active use by a large number of scientists already. Toolkit routines have also been incorporated as parts of other, non-portable, programs on a variety of platforms.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00033-02-IEB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Unification of Biotechnology Information

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

James M. Ostell, Chief, IEB, NCBI, NLM
Mark Cavanaugh, Computer Specialist, NCBI, NLM
Karl Sirotkin, Research Biologist, NCBI, NLM
Carolyn Tolstoshev, Visiting Associate, NCBI, NLM

COOPERATING UNITS *(If any)*

LAB/BRANCH

Information Engineering Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
4	4	0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

To provide a single formal specification for information relevant to biotechnology computing, including scientific literature, nucleic acid sequence data, protein sequence data, genetic and physical maps, chromosomes, genes, the relationships of other scientific knowledge about these entities and their relationship to normal and disease conditions. To convert a number of important biological databases of diverse content and form into such unifying specification. Develop tools demonstrating use of such unified view of biological data. A large number of databases were examined, such as GenBank, EMBL, PIR, SWISSPROT, Kabat, MIM, ACEDB, Flybase, EcoSeq, MEDLINE, and others. A single modular data model was constructed which could represent almost all of the data from these sources in a consistent way, and formally specified in Abstract Syntax Notation 1, (ISO 8824, 8825). Parsers were written to read the different data formats of the sources, and software developed to map the different available data elements into the proper places in the unified ASN.1 data model. A software product (Entrez) was developed to production quality which took advantage of the unified view of some of the sources (GenBank, PIR, SWISSPROT, and MEDLINE) to allow the scientist to explore all these data as a single integrated whole. Entrez and it's associated integrated data is distributed to scientists on CDROM every two months by NCBI. A client/server version provides high speed access over Internet. New databases now allow NCBI to maintain an "up to the minute" view of the diverse data sources and their relationships with each other despite differing content, data formats, and data release cycles. Additional data sources, such as 3-D protein structures, are being mapping to the common specification and new tools are being developed to use the growing web of connected data. Scientific knowledge continues to evolve, which means the data must evolve as well. This is a project which must continue as long as biomedical research does.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER: Z01-LM-00034-02-BRB

PERIOD COVERED
October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*
3-D Computer Modeling of Ribonucleic Acids

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*
Francois Major, Visiting Fellow, BRB, NCBI, NLM
R. Cedergren, Professor, University of Montreal

COOPERATING UNITS *(if any)*
Department of Biochemistry, University of Montreal

LAB/BRANCH
Basic Research Branch

SECTION

INSTITUTE AND LOCATION
NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.8	0.8	0.0

CHECK APPROPRIATE BOX(ES)
☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

A. We developed and tested a protocol to reduce subjectivity from the process of ribonucleic acid (RNA) modeling. B. The modeling of specific RNA molecules has been continued.

A. We have developed and extensively tested a modeling protocol to be applied with the computer program MC-SYM. The protocol is based on simple RNA structural principles such as base stacking and pairing. The protocol consists of a systematic search over the space of input scripts for the most restrictive one that produces solutions. This strategy allows to keep the number of solutions low and therefore simplifies a posteriori evaluation. The modeling of a transfer RNA (tRNA) molecule from structural data that was available prior to the crystal structure was used to evaluate the protocol quantitatively; by measuring the quality of the solutions produced by root mean square (rms) deviation from known crystal structure of tRNA-Phe and tRNA-Asp. The application of this protocol to tRNA allowed for the generation of the best model in almost all cases.

B. The specific modeling of RNA molecules of biological interest is still in progress. These molecules are: a small metalloribozyme called leadzyme, group I introns, Ribonuclease P RNAs, and a spliceozyme. The leadzyme allowed us to study non-canonical base pairings and nucleotide binding to ions such as Pb2+ and Mg2+. The other molecules allow us to study the relative orientation and position of helices in large RNAs.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00035-02-BRB

PERIOD COVERED

November 15, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Genome organization and evolution of RNA viruses

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Eugene V. Koonin Visiting Scientist, NCBI, NLM
V. V. Dolja, Biology Dept., Texas A&M Unviersity

COOPERATING UNITS *(if any)*

Biology Department, Texas A&M Unviersity, College Station, TX

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.2	0.2	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

Detailed computer-assisted comparative analysis of protein sequences and gene arrangement in positive-strand and double-stranded RNA viruses was performed in order to derive principles of genome organization, and to propose evolutionary scenarios and to suggest a new, phylogeny-based taxonomy.

Despite the rapid mutational change that is typical of positive-strand RNA viruses, enzymes mediating the replication and expression of virus genomes contain arrays of conserved sequence motifs. Proteins with such motifs include RNA-dependent RNA polymerase, putative RNA helicase, chymotrypsin-like and papain-like proteases, and methyltransferases. The genes for these proteins form partially conserved modules in large subsets of viruses. A concept of the virus genome as a relatively evolutionarily stable "core" of house-keeping genes accompanied by a much more flexible "shell" consisting mostly of genes coding for virion components and various accessory proteins is discussed. Shuffling of the "shell" genes including genome reorganization and recombination between remote groups of viruses is considered to be one of the major factors of virus evolution.

Multiple alignments for the conserved viral proteins were constructed and used to generate the respective phylogenetic trees. Based primarily on the tentative phylogeny for the RNA-dependent RNA polymerase, which is the only universally conserved protein of positive-strand RNA viruses, three large classes of viruses, each consisting of distinct smaller divisions, were delineated. Strong correlation was observed between this grouping and the tentative phylogenies for the other conserved proteins as well as the arrangement of genes encoding these proteins in the virus genome. A comparable correlation with the polymerase phylogeny was not found for genes encoding virion components or for genome expression strategies. It is surmised that several types of arrangement of the "shell" genes as well as basic mechanisms of expression could have evolved independently in different evolutionary lineages.

The grouping revealed by phylogenetic analysis may provide the basis for revision of virus classification, and phylogenetic taxonomy of positive-strand RNA viruses is outlined. Some of the phylogenetically derived divisions of positive-strand RNA viruses include also double-stranded RNA viruses indicating that in certain cases the type of genome nucleic acid may not be a reliable taxonomic criterion for viruses.

Hypothetical evolutionary scenarios for positive-strand RNA viruses were proposed. It is hypothesized that all positive-strand RNA viruses and some related double-stranded RNA viruses could have evolved from a common ancestor virus that contained genes for RNA-dependent RNA polymerase, a chymotrypsin-related protease that functioned also as the capsid protein, and possibly an RNA helicase.

The significance of the project lies in the unique opportunity to compare the organization of a large number of complete genomes with varying degree of similarity to each other, to derive general principles of generation of most parsimonious evolutionary scenarios, and to propose new, phylogenetict axonomy of RNA viruses.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00037-02-BRB

PERIOD COVERED

September 30, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Analysis of conserved amino acid sequence motifs in NTPases.

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Eugene V. Koonin Visiting Scientist, NCBI, NLM

COOPERATING UNITS *(if any)*

Institute of Poliomyelitis and Viral Encephalitides, Russian Academy of Medical Sciences, Moscow, Russia (A. E. Gorbalenya)

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.2	0.2	0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☑ (c) Neither

☐ (a1) Minors

☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

Conserved amino acid sequence motifs in different groups of NTP-utilizing enzymes were studied using computer methods of sequence analysis to the end of predicting NTPase activity of unexplored proteins, designing schemes for identification of NTPases in sequence databases and generating a sequence-based classification of this type of enzymes.

NTPases are characterized by well-defined conserved motifs that are implicated in substrate binding and hydrolysis. Amino acid sequence databases were searched for the so-called A motif that is involved in phosphate binding and the resulting set of proteins was explored in detail with respect to their similarities with other proteins, and the available data on NTPase activity. A new superfamily of (putative) DNA-dependent ATPases was described that includes the ATPase domains of prokaryotic NtrC-related transcription regulators, MCM proteins involved in the initiation of eukaryotic DNA replication, and a group of uncharacterized bacterial and chloroplast proteins. MCM proteins were shown to contain a modified form of the ATP-binding motif and are predicted to mediate ATP-dependent opening of double-stranded DNA in the replication origins. In a second line of investigation, it was demonstrated that the products of unidentified open reading frames from Marchantia mitochondria and from yeast, and a domain of a baculovirus protein involved in viral DNA replication are related to the superfamily III of DNA and RNA helicases that previously has been known to include only proteins of small viruses. Comparison of the multiple alignments showed that the proteins of the NtrC superfamily and the helicases of superfamily III share three related sequence motifs tightly packed in the ATPase domain that consists of 100 - 150 amino acid residues. A similar array of conserved motifs was found in the family of DnaA-related ATPases. It is hypothesized that the three large groups of nucleic acid-dependent ATPases have similar structure of the core ATPase domain and have evolved from a common ancestor. Several previously uncharacterized proteins were shown to contain conserved sequence motifs typical of the helicase superfamilies I or II and were predicted to possess helicase activity. A general classification of DNA and RNA helicases based on sequence comparison was outlined and an attempt was made to derive identifying sequence pattern for each large group.

The significance of the project is in the prediction of NTPase activity for many proteins with unknown functions, characterization of allowed deviations in NTP-binding motifs, derivation of identifying patterns for different groups of NTPases, and development of a sequence-based classification for a vast enzyme class.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00038-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

The Cold Shock Domain (CSD) Protein Motif.

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

David Landsman, Visiting Scientist, BRB, NCBI, NLM
S.H. Bryant, Research Biologist, BRB, NCBI, NLM
J. Doniger, Georgetown Unversity, Washington DC.
M.A. Gonda, LCMS, NCI-Frederick
G. Wistow, Visiting Scientist, LMDB, NEI.

COOPERATING UNITS *(If any)*

Laboratory of Molecular and Developemnet Biology, National Eye Institute, National Institutes of Health, Bethesda, MD (G. Wistow).
Laboratory of Cell and Molecular NCI-Frederick Cancer Research and Developement Center, Frederick, MD (M.A. Gonda).
Georgetown Unversity Medical Center, Washington, DC (J. Doniger).

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
1.5	1.5	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

The open reading frame transcribed from the *unr* gene (immediately upstream of N-ras) in mammals consists of multiple repeats similar to the cold-shock domain, a putative DNA-binding motif found in prokaryotic cold-shock proteins and eukaryotic DNA-binding proteins. Alignment of the CSD sequences of *unr* with those from other proteins reveals a core of similarity for which a consistent secondary structure prediction can be derived. This prediction suggests that the CSD consists primarily of β-sheet, in contrast to most known eukaryotic DNA-binding proteins. Sequence analysis of the 3′ end of the guinea pig *unr* gene shows that the core of one CSD repeat is encoded in a single exon, consistent with the modular assembly of the gene from ancestral CSD-coding units (Doniger et al. 1992). Further sequence analysis has shown that there is a short motif of 8 amino acids, corresponding to the RNP-1 motif found in canonical RNA-binding domains (Landsman, 1992). The CSD family of proteins, which includes several transcription factors which have been shown to bind specifically to DNA, has now been identified to contain a motif similar to RNP-1. A non-redundant protein sequence database was searched with regular expressions and with a weight/residue position matrix of the RNP-1 motif resulting in the identification of numerous known members of the RNA-binding family of proteins. In addition, the search identified that the CSD-containing family of proteins includes a motif which is almost identical to the RNP-1 motif. A determination of the statistical significance of this analysis showed that the RNP-1 motifs from these two families of proteins are indeed similar. It is conceivable that the RNP-1 in the CSD-containing proteins enables them to function as both double- and single-stranded, DNA- and RNA-binding proteins. This suggests that the CSD-containing proteins could be involved in transcription as well as in gene regulation post-transcriptionally by binding RNA. The initial phases of modelling a CSD based on the crystal structure of an RNA-binding protein have been attempted.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00041-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Informatics Analysis of the E. coli Genome

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Kenneth E. Rudd, Research Biologist, NCBI, NLM
Jim Ostell, Chief, IEB, NCBI, NLM
Gerald Bouffard, Graduate Student, George Washington Univ., Washington, D. C.
Carolyn Tolstoshev, Visiting Scientist, IEB, NCBI, NLM
Jinghui Zhang, Graduate Student, NCBI, NLM

COOPERATING UNITS *(if any)*

Department of Computer Science, The Penn State University, University Park, PA.
Department of Microbiology, George Washington University, Washington, D. C.
Biological Science Group, University of Connecticut, Storrs, CT
Dept. of Medicine, Mt. Sinai Medical Center, New York, NY

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.7	0.7	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

The E. coli genome contains over 3000 genes and is currently over 50% sequenced. A complete high resolution restriction map for the entire genome is available. This makes the E. coli genome project the most advanced of all cellular genome projects. This information has been collected and organized into a cohesive information base, unifying the efforts of many laboratories into a single data resource. This project includes software development, database development, and data analysis. The software that has been developed or enhanced during the reporting period include a new genomic sequence viewer and editor, ChromoScope. Two relational databases have been developed: GeneScape, a Macintosh database of genomic map information that is essentially completed, and EC-BASE, a Sybase database of E. coli map and DNA sequence information. Report generators for ECBASE now allow complete ASN.1 and GenBank flatfile reports of EcoSeq contigs, DNA sequence and genomic restriction map data has been analyzed to determine the information content of ribosome binding sites, number and distribution of genomic restriction sites, repeated patterns in DNA sequences, distribution and categorization of proteins encoded in the genome, assignment of genes to individual clones in the ordered clone set of the E. coli genome, and the detection of putative new genes in the DNA sequence flanking known genes. One sequencing gap has been closed, revealing the sequence of two new E. coli genes, gpmA and galM.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00042-02-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Extended Two Dimensional Lattice Models of Proteins

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and Institute affiliation)*

David J. Lipman, Director, NCBI, NLM
W. John Wilbur, Senior Scientist, BRB, NCBI, NLM

COOPERATING UNITS *(If any)*

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.1	0.1	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

Extensions of two dimensional lattice models of proteins have been developed to better understand their evolution and general biophysical properties.

Previous highly idealized lattice models for proteins developed by Dill have been shown to display many of the general properties of proteins. Previous work by us showed that these models also display evolutionary behavior which parallels that of biological sequences. The current work generalizes these lattice models to larger alphabets (i.e., more different classes of residues), and subsequently more complex interaction rules.

Evaluating this system using the previous criteria for folders indicates that such sequences have higher probabilities of folding to unique structures, and more different structures are obtained. This is in the context of effectively zero temperature. Moving to nonzero temperatures increases the realism of the model and an information theoretic framework is developed which evaluates the specificity of sequence for structure. In addition, a simple probabilistic model is developed which provides a basis for generating optimal interaction rules with respect to maximizing the information of the sequence-to-structure mapping.

It is found that for realistic temperatures and low average interaction energies (in the range found in proteins), a single binary interaction corresponding to domination by the hydrophobic effect is optimal.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00045-01-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Structure Prediction by Protein Threading

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Stephen H. Bryant, Senior Investigator, BRB,NCBI, NLM
Charles E. Lawrence, Biometrics Laboratory, WCLR, New York State Department of Health

COOPERATING UNITS *(if any)*

Biometrics Laboratory, Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, NY

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.5	0.5	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

We have developed computer methods to predict protein three-dimensional structure by recognition of folding motif. A protein's sequence is "threaded" through alternative backbone structures in a database, and the conformations most stabilized by that sequence are identified. The work has focused on three areas: 1) derivation of a rapidly-evaluated empirical free energy function, 2) testing of algorithms for fast threading through gapped "core" motifs, and 3) identification of core substructures to be included in the database. This year we have shown that an energy function based on residue contact potentials can identify the correct core motif and alignment among many billions of alternatives, a specificity sufficient for many prediction problems. We have developed statistical corrections for effects of sequence composition and length, which otherwise reduce motif-recognition specificity. We have tested a fast-threading algorithm based on a monte carlo technique, and found that it is capable of identifying optimal alignments of sequence and fixed-size core motifs in a few seconds, in all cases where this optimum can be determined with certainty by enumeration. To define a database of folding motifs we have extracted from the Protein Data Bank substructures consisting of major helices and beta-strands, using an algorithm resistant to local distortions and/or coordinate imprecision. Each defines a family of related core motifs, when threaded with a monte carlo algorithm that allows deletions and/or changes in size of individual secondary structure elements. Testing of this adaptive threading algorithm and its associated database is in progress. The significance of this research is that these methods may allow automated search of a folding motif database, predicting substructure conformations in proteins which may share little or no homology with proteins in the crystallographic database.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00046-01-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Databases for Molecular Modeling

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Stephen H. Bryant, Senior Investigator, BRB,NCBI, NLM
James Ostell, Senior Investigator, SEB, NCBI, NLM
Hitomi Ohkawa, Postdoctoral Fellow, BRB, NCBI, NLM

COOPERATING UNITS *(if any)*

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.95	0.95	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither

☐ (a1) Minors

☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

We have developed databases and software useful for modeling of protein three-dimensional structure and analysis of sequence-structure relationships. These tools have been distributed freely to biologists and developers of biotechnology software. The work may be divided into three areas: 1) continued development of the "PKB" object-oriented research database, 2) production of GenBank entries describing proteins and nucleic acids with known structure, and 3) design and implementation of a structural database convenient for developers of molecular modeling software. The PKB data specification has been expanded to include all data items defined by the Protein Data Bank, including scientific literature citations and bonded connectivity. We have also added validation procedures which recover correct definitions of biopolymer sequence and chemical modification, and identify stereochemical anomolies. We have used PKB to produce ASN.1-language reports of structural features mapable to sequence, using the current NCBI/GenBank data specification. These data have been updated as new structures became available from the Protein Data Bank, and incorporated into the widely distributed GenBank/ASN.1 and Entrez databases. To provide convenient access to structural data from modern application programs written in C we have begun development of an ASN.1 database containing complete covalent and spatial structure data. Its specification allows comparison of biopolymer or non-biopolymer components of biological macromolecules according to chemical structure, and direct representation of three-dimensional structure inferred by alignment with homologous or chemically similar molecules. The significance of this work is in providing biologists with easy access to structural data, and in providing a software infrastructure to researchers interested in sequence-structure relationships and programmers developing molecular modeling software.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00047-01-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

A representative set of protein sequences for similarity

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Stephen Altschul, Senior Staff Fellow, NCBI, NLM, NIH
Warren Gish, Staff Fellow, NCBI, NLM, NIH
Todd Lowe, Computer Scientist, NCBI, NLM, NIH
David J. Lipman, Director, NCBI, NLM, NIH

COOPERATING UNITS *(if any)*

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.4	0.4	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

Recent genomic data have revealed that approximately two fifths of all cellular proteins contain regions that have been detectably conserved among species evolutionarily diverged by over 500 million years. It is estimated that fewer than 1000 such regions exist, and that current protein databases contain most of them. Proteins containing these regions will be of particular use in identifying novel sequences from the various genome projects. We have therefore attempted to construct a relatively small "core" database of such proteins. This database, while under 15% the size of comprehensive proteins sequence databases, is nevertheless able to detect virtually all the significant similarities that newly sequenced proteins show to the complete databases. It may be used to speed up database searches and to reduce the amount of redundant output they generate, thereby improving as well the ability to detect relatively subtle similarities. The core database is also interesting as an object of study in its own right, permitting improved statistical studies of protein sequences and suggesting the broad outline of proteins necessary for life. Our aim has not been to provide a rigorous definition for the core of a protein sequence database, but rather to construct a simple but useful tool where none existed before.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00048-01-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Identification and mapping of the human homolog of the yeast cell cycle gene

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Mark S. Boguski, Senior Medical Staff Fellow, BRB, NCBI, NLM
Philip Hieter, Assoc. Prof., Johns Hopkins University, School of Medicine, Baltimore, MD

COOPERATING UNITS *(if any)*

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.2	0.2	0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

Previous work has shown that CDC27 acts late in G2 of the cell cycle, after DNA replication but prior to the onset of chromosome segregation and that the encoded protein binds to the mitotic spindle. Several unsuccessful attempts were made to clone the human homolog by conventional techniques. A search of dbEST (see report number Z01-LM-00015-01-BRB) showed a weak but significant match to a human brain protein. Additional sequencing confirmed the homology and the DNA was mapped to human chromosome 17q21-24 thus becoming a candidate gene for human breast cancer susceptibility. A computer study was performed using a training set of known yeast-human homolog pairs to assess the general utility of this method of gene discovery and to optimize search parameters.

Over the next three years we will systematically search for new yeast-human homolog pairs and map approximately 1000 of these to human chromosomes. We expect to find a number of yeast proteins that will become candidates for human disease genes by virtue of their map locations.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER: Z01-LM-00050-01-BRB

PERIOD COVERED
October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*
DNA sequence complexity and mutational dynamics.

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*
John Wootton, Visiting Scientist, BRB, NCBI, NLM

COOPERATING UNITS *(if any)*
P. Salamon & J.D. Nulton, Dept. of Mathematical Sciences, San Diego State Univ;
A.K. Konopka, NCI/DCBRC, Frederick;
L.K. Hansen, CONNECT Electronics Institute, Technical University of Denmark.

LAB/BRANCH
Basic Research Branch

SECTION

INSTITUTE AND LOCATION
NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.3	0.3	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☑ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

DNA sequences show enourmous statistical heterogeneity in their local compositional characteristics. This project is investigating the possible causes of the mosaicism in sequence complexity, particularly the roles of mutational biases, medium and long range combinatorial constraints, and correlates of specific functional and structural classes of sequence.

A. Distributions of local complexity: Using formal definitions of local compositional complexity of DNA subsequences, the robustness of the Salamon/Konopka maximum entropy relationship has been explored: Given a functionally equivalent set of DNA sequences, the distribution of complexity among all subsequences of this set appears to be as random as possible consistent with the mean complexity of these subsequences. It has now been shown that this maximum entropy effect follows as a consequence of the dynamics of almost any mutational mechanism that incorporates a bias toward low-complexity, for example the neighbor-dependent biases in substitution mutations observed in human genomic mutations.

B. Feathered structure of medium range complexity distributions: DNA segments of length range 40 to 200 nucleotides have distributions of compositional complexity with a novel regularity of structure ("feathering"). This is not yet fully explained but depends partly on combinatorial constraints and partly on the nonuniform representation of different complexity classes in natural DNA sequences.

Significance of project: The project is beginning to provide detailed explanations for some of the puzzling features found from statistical analyses of DNA sequences, including aspects of the so-called "long range correlations" inferred by other research groups from spectral analysis.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER: Z01-LM-00051-01-BRB

PERIOD COVERED
October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*
Subtle sequence patterns in DNA-binding complexes

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*
John Wootton, Visiting Scientist, BRB, NCBI, NLM

COOPERATING UNITS *(if any)*
M. Protic & A.S. Levine, NICHD/NIH;
Y. Nakatani & D-W. Gong, NINDS & NICHD/NIH

LAB/BRANCH
Basic Research Branch

SECTION

INSTITUTE AND LOCATION
NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.2	0.2	0.0

CHECK APPROPRIATE BOX(ES)
☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

The number of classes of proteins that recognize specific sequences and structures of DNA is continuing to grow rapidly. Several are large complexes of several types of subunit. In this project, computational strategies are being developed to recognize some of the functional amino acid sequence patterns involved, many of which are subtle and variable. Improved methods are being applied to the analysis of new sequences determined in other NIH laboratories.

For patterns of well-established DNA binding regions such as helix-turn-helix and basic regions, new methods for evaluation of the diagnostic power of pattern discriminators are being developed and applied to more novel patterns. Low- complexity sequences are frequent in DNA-binding proteins and require analysis and filtering before database searches. The new method of automated local multiple alignment by iterative sampling is also included in these strategies.

Specific sequences analyzed include (1) 127 kDa component of a UV-damaged-DNA binding complex, which is defective in some Xeroderma Pigmentosum Group E patients and shows sequence similarities to proteins of unknown function from a wide range of eukaryotes, and (2) six subunits of the transcription factor TFIID complex that is central in transcriptional regulation by facilitating promoter responses to various activators.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00052-01-BRB

PERIOD COVERED

November 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

A Depth-First Search Algorithm for Detecting Patterns in Protein Sequences.

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Andrew F. Neuwald, IRTA Fellow, NCBI/NLM
Philip Green, Dept of Genetics, Washington Univ School of Medicine, St. Louis, MO

COOPERATING UNITS *(if any)*

Department of Genetics, Washington University School of Medicine, St. Louis, MO

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.35	0.35	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

This project continues work on a method for detecting motifs in protein sequences that was started at Washington University in St. Louis. The basic method, which uses a depth-first strategy to search for statistically significant patterns, was extended by A. enhancing the basic algorithm and B. incorporating additional procedures to process the output. A. Enhancing the basic algorithm. First, the speed of the basic depth-first search algorithm was increased by a factor of about ten. Second, detection of subtle motifs was enhanced by including patterns having pairs of related residues in the search (the original method searched only for single residue patterns). (Another modification, which allows incorporation of a similarity scoring matrix, was also developed.) Third, the statistical method for estimating pattern p-values was improved. B. Incorporating additional procedures. Two new procedures were developed that take as input the output from the depth-first search algorithm and attempt to identify protein regions sharing a common motif. Since a search can yield a large number of related patterns the first procedure groups these patterns and identifies the matching regions. Some regions, however, may share significant similarity to a motif without matching a statistically significant pattern and conversely, regions that otherwise have no significant similarity to a motif may match a pattern by chance. Therefore the second procedure attempts to correct for this in order to identify those regions most likely to share a common motif. The significance of the methods developed during this project lie in their ability to find subtle but significant motifs that are not detectable by other currently available methods. Motifs usually correspond to structurally and functionally conserved regions so that their detection can aid the experimentalist in protein characterization and classification. A manuscript describing these methods and their application to several very difficult problems will soon be submitted for publication.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER: Z01-LM-00053-01-BRB

PERIOD COVERED
August 1, 1993 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*
Situation Theory as a Model for Ontological Engineering and Knowledge

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*
John Wilbur, Senior Scientist, BRB, NCBI, NLM
Karl Sirotkin, Software Engineer, NCBI, NLM, NIH

COOPERATING UNITS *(If any)*

LAB/BRANCH
Basic Research Branch

SECTION

INSTITUTE AND LOCATION
NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.2	0.2	0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

We believe that situation theory as described by Devlin in his recent book, *Logic and Information* (1991) may provide a useful approach to ontological engineering. Situations provide a flexible method of knowledge representation well suited to take advantage of a light weight parse of natural language into phrase units. With this approach it is our purpose to design a system that takes advantage of as much automatic processing as possible to aid in knowledge acquisition but which relies ultimately on the human to correct the process. As the knowledge base grows it will in turn be used to aid in the process of acquisition. As conceived the initial stages will be the most labor intensive. The first task for the system will be to test its usefulness on a small set of documents to determine how it may aid retrieval. If results are positive we may proceed to a more ambitious effort to build a large base of information in molecular biology.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00054-01-BRB

PERIOD COVERED

September 30, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Comprehensive computer analysis of E. coli genes.

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Eugene V. Koonin Visiting Scientist, NCBI, NLM
Kenneth E. Rudd, Research Biologist, NCBI, NLM
R. L. Tatusov, Visiting Fellow, NCBI, NLM
M. Borodovsky, Department of Biology, Georgia Institute of Technology, Atlanta, GA

COOPERATING UNITS *(if any)*

Department of Biology, Georgia Institute of Technology, Atlanta, GA (M. Borodovsky)

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.3	0.3	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

With the fraction of known sequences in the Escherichia coli chromosome nowexceeding 50 per cent, the goal of comprehensive computer analysis of the bacterial genome is becoming realistic. The scope of this project includes development of an optimal strategy for analysis of the genetic contents of the genome; assessment of the utility of different computer-assisted methods in large-scale genome projects; identification of all genes in the bacterial chromosome; and extraction of maximal amount of information on possible functions and evolutionary relationships of gene products; delineation off possible regularities in the distribution of related genes in the bacterial chromosome. Comparison of the 1400 protein sequences contained in the EcoSeq6 database with the complete amino acid sequence databases was performed, with particular emphasis on the relationship between various E.coli proteins. A variety of computer methods for database search, motif identification, and multiple sequence alignment were employed, including newly developed algorithms. As the result, probable functions were predicted for a number of previously uncharacterized putative open reading frame products, and several new proteins families and highly conserved, probably functionally important sequence motifs were described. The most interesting findings included: a putative new system of regulated, GTP-dependent proteolysis; a family of putative GTP phosphohydrolases related to the antimutator protein MutT, with an apparent GTP-binding motif of a novel type; two previously uncharacterized DNA or RNA helicases belonging to distinct groups within the "DEAD/H" superfamily; several unknown putative methyltransferases. New, unexpected relationships were found for proteins that have been previously characterized functionally, but not structurally, e.g. it was shown that diadenosine tetraphosphate phosphohydrolase (ApaH) is related to protein phosphatases; and RNase T is related to DNA proofreading exonucleases. Regions of the E.coli chromosome that have been annotated as untranslated in the EcoSeq6 database were explored using the GENMARK method for coding region prediction and BLASTX program for database search. As the result, about 100 new genes were predicted to exist in the E.coli chromosome encoding putative enzymes, membrane proteins, and regulatory proteins. Strong correlation was established between the results of GENMARK prediction and similarity search, suggesting that the coding regions predicted by GENMARK, but not showing similarity to
sequences available in current databases are still likely to correspond to new genes.

The significance of the project lies in the potential for development of optimal strategy for computer analysis of gene functions and arrangement at the whole genome scale; and in the prediction of likely functions for many gene products leading to stimulation of further experimental dissection.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER: Z01-LM-00055-01-BRB

PERIOD COVERED
October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*
The HMG-1 Box Protein Motif

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

David Landsman, Visiting Scientist, BRB, NCBI, NLM
S.H. Bryant, Research Biologist, BRB, NCBI, NLM
A.D. Baxevanis, NRC Fellow, BRB, NCBI, NLM
M. Bustin, Research Biologist, LMC, DCE, NCI

COOPERATING UNITS *(If any)*

LAB/BRANCH
Basic Research Branch

SECTION

INSTITUTE AND LOCATION
NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.0	0.0	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

In the past few years, numerous proteins have been identified as containing a stretch of about 75 amino acids which are homologous to an abundant non-histone chromosomal protein HMG-1. These proteins bind DNA and bends it on binding or bind preferentially bind to bent DNA. Several of these proteins have been implicated in numerous nuclear functions including transcription, replication, and chromatin structure as well as transcription regulation in mitochondria resulting, in some cases, in such phenotypes as sex and mating type determination. Examples of these proteins are UBF, an RNA polymerase I transcription factor; SRY, the mammalian testis-determining factor; LEF-1/TCF1a, the lymphoid enhancer binding factor; ABF2 and mtTF1, mitochondrial transcription factors; and T160, a V-(D)-J recombinase which is very similar to SSRP1, a structure specific recognition protein that binds cisplatin-modified DNA. In addition, there are several yeast proteins involved in mating type determinations and sexual development. We are compiling and maintaining a database of the HMG-1 box family of proteins and are analyzing the sequences to determine the phylogeny between these functionally widely-diverse proteins. Recently, the 3D structure of an HMG-1 box from rat HMG-1 has been determined by two dimensional NMR spectroscopy (Weir, et al. 1993). We plan to use this structure to improve our alignment and to model other members of the family.

DEPARTMENT OF HEALTH AND HUMAN SERVICES - PUBLIC HEALTH SERVICE

NOTICE OF INTRAMURAL RESEARCH PROJECT

PROJECT NUMBER

Z01-LM-00056-01-BRB

PERIOD COVERED

October 1, 1992 to September 30, 1993

TITLE OF PROJECT *(80 characters or less. Title must fit on one line between the borders.)*

Gibbs sampling methods for the analysis of biopolymer sequence data.

PRINCIPAL INVESTIGATOR *(List other professional personnel below the Principal Investigator.) (Name, title, laboratory, and institute affiliation)*

Charles Lawrence, IPA, NCBI, NLM
Stephen Altschul, Senior Staff Fellow, NCBI, NLM
Mark Boguski, Senior Medical Staff Fellow, NCBI, NLM
Andrew Neuwald, IRTA Fellow, NCBI, NLM
John Wootton, Visiting Scientist, NCBI, NLM

COOPERATING UNITS *(If any)*

Wadsworth Center for Laboratories and Research, NYS-DOH, Albany, NY
Statistics Department, Harvard University

LAB/BRANCH

Basic Research Branch

SECTION

INSTITUTE AND LOCATION

NCBI, NLM, Bethesda, MD 20894

TOTAL STAFF YEARS:	PROFESSIONAL:	OTHER:
0.75	0.75	0.0

CHECK APPROPRIATE BOX(ES)

☐ (a) Human subjects ☐ (b) Human tissues ☒ (c) Neither
☐ (a1) Minors
☐ (a2) Interviews

SUMMARY OF WORK *(Use standard unreduced type. Do not exceed the space provided.)*

Multiple sequence alignment has proved to be a remarkably successful means of representing and organizing much of the present deluge of inferred protein sequence data. It is crucial to research on the structure and function of proteins, promoting the detection and description of sequence motifs and aiding efforts at protein modeling, structure prediction and engineering. In addition, by organizing information on mutational variation, multiple sequence alignment can elucidate molecular evolution and serve as the input for phylogenetic reconstruction. The importance of local multiple sequence alignment has long been appreciated and has been the subject of extensive study. The goal of automated methods is to produce optimized alignments, using only the information intrinsic to the sequences themselves. Unfortunately, rigorous algorithms for finding optimal solutions are so computationally expensive as to limit their application to a very small number of sequences. On the other hand, many heuristic approaches gain speed at the sacrifice of sensitivity to subtle patterns. We have developed a new statistically based algorithm that aligns sequences by means of predictive inference. Using residue frequencies, this Gibbs sampling algorithm iteratively selects alignments in accordance with their conditional probabilities. The newly formed alignments in turn update an evolving residue frequency model. When equilibrium is reached the most probable alignment can be identified. If a detectable pattern is present, we have found convergence is rapid. Effectively, the algorithm finds optimal local alignments from multiple sequences in linear time (seconds on current workstations). Its use is illustrated on test sets of helix turn helix proteins, lipocalins, and prenyltranferases. Continuing work on this project focuses on the relaxation of assumptions concerning motif size and number, and the independence of input sequences. The inclusion of residue interactions will also be explored.